10-2010

Bugs on a Budget: Distributed Sensing With Cost for Reporting and Non-Reporting

Vladamir Pozdnyakov
University of Connecticut

J Michael Steele
University of Pennsylvania

Recommended Citation

Pozdnyakov, V., & Steele, J. M. (2010). Bugs on a Budget: Distributed Sensing With Cost for Reporting and Non-Reporting. *Probability in the Engineering and Informational Sciences, 24* (4), 525-534. http://dx.doi.org/10.1017/S026996481000015X

Bugs on a Budget: Distributed Sensing With Cost for Reporting and Non-Reporting

Abstract
We consider a simple model of sequential decisions made by a fusion agent that receives binary-passive reports from distributed sensors. The main result is an explicit formula for the probability of making a decision before a fixed budget is exhausted. These results depend on the relationship between a special ruin problem for a "lazy random walk" and a traditional biased walk.

Disciplines
Business | Statistics and Probability

Probability in the Engineering and Informational Sciences, **24**, 2010, 525–534.
doi:10.1017/S026996481000015X

BUGS ON A BUDGET

DISTRIBUTED SENSING WITH COST FOR REPORTING AND NONREPORTING

VLADIMIR POZDNYAKOV
Department of Statistics
University of Connecticut
Storrs, CT 06269-4120
E-mail: vladimir.pozdnyakov@uconn.edu

J. MICHAEL STEELE
Department of Statistics
University of Pennsylvania
Philadelphia, PA 19104
E-mail: steele@wharton.upenn.edu

We consider a simple model of sequential decisions made by a fusion agent that receives binary-passive reports from distributed sensors. The main result is an explicit formula for the probability of making a decision before a fixed budget is exhausted. These results depend on the relationship between a special ruin problem for a "lazy random walk" and a traditional biased walk.

1. DISTRIBUTED SENSING WITH DECENTRALIZATION AND FUSION

For many tasks of reconnaissance and surveillance, networks of spatially distributed sensors provide the low-cost, low-risk solution, and the rapidly growing field of distributed sensing now provides many interesting challenges for probabilistic modeling (see, e.g., Chong and Kumar [1]). The main purpose of this article is to examine a simple model that joins the rudiments of sequential decision-making with the engineering constraint of a fixed budget for the cost of the transmissions sent to and from the distributed sensors. Here the costs associated with transmissions from the sensors are intended to either real physical expenditures, such as battery life, or to capture

 0269-9648/10 $25.00

more subtle costs, such as the cumulative risk of a remote sensor (or bug) being found by an adversary.

Our focus is specifically on distributed sensor networks with a *decentralized architecture*; that is, each sensor is capable of certain preliminary decision processing before sending summarized information to an automated *fusion agent*. The fusion agent is thought of as being remote from the distributed sensors, and it has two responsibilities. First, it interrogates the distributed sensors according to some protocol and, second, it makes an "overall" network decision at such as time as sufficient evidence has been collected to support a decision. In the protocols considered here, the fusion agent remotely interrogates the distributed sensors sequentially one by one.

We are mostly concerned with a "binary-plus-passive" design for which at the time a sensor is interrogated, it checks its locally stored observation data and acts according to a three-part rule: (1) If the observation data has a "large weight" (in some appropriate sense), it sends $+1$ to the fusion center; (2) If it has a correspondingly "small weight", then the sensor sends -1; and, (3) if the observation is within a certain intermediate range, the sensor *does not reply at all*. Here the fusion agent *does* know that the sensor was interrogated, so the nonreply conveys useful information. The benefit of this "passive reply" alternative is that the sensor does not expend energy or expose itself to incremental risk of adversary detection. Naturally, we view an active response from a sensor as expensive and we view a nonresponse as relatively inexpensive—but not entirely costless.

2. TRACKING THE COSTS AND THE TIME TO DECISION

To fix notation, we first describe a version of our problem that is a bit more general than the one we analyze in detail. By $\{X_i\}_{i\geq 1}$ we denote a sequence of discrete-valued, independent and identically distributed (i.i.d) random variables that we view as "information weights" coming from of the reports (or nonreports) from a sequence of queries that are put to the distributed sensors. Next, we consider a nonnegative function $c(\cdot)$ and we view $c(X_i)$ as the cost of the fusion agent collecting the value X_i, either directly from the sensor report or through the information of a "nonreport." We then let B denote our budget for payment for the costs of this collected information.

We assume that we can continue to collect distributed sensor information so long as our cumulative cost C_n satisfies the budget constraint

$$C_n \stackrel{\text{def}}{=} c(X_1) + c(X_2) + \cdots + c(X_n) < B,$$

and we assume that decision rule for the fusion agent is based on the sum of the signal information weights

$$S_n = X_1 + X_2 + \cdots + X_n.$$

A decision is made by the fusion agent at the time when the process $\{S_n\}$ hits either an upper boundary $\mathcal{U} \equiv \{U(n) : n = 1, 2, \ldots\}$, say for a positive decision, or a lower boundary $\mathcal{L} \equiv \{L(n) : n = 1, 2, \ldots\}$ for a negative decision. Naturally, the upper and

lower boundaries depend on the hypothesis that the fusion must test to make the network decision. In general, these boundaries are curved, and they would be determined by the usual tools of sequential decisions theory such as the sequential probability ratio test (or its extensions and approximations).

Thus, our distributed sensing problem leads to a special kind of boundary crossing problem for the two-dimensional process $Z_n = (C_n, S_n), n = 1, 2, \ldots$. From the design prospective, the random variables of most interest are the *decision time*,

$$\tau_D = \min\{n : S_n \geq U(n) \text{ or } S_n \leq L(n)\}, \tag{1}$$

and the *budget exhaustion time*,

$$\tau_E = \min\{n : C_n \geq B\}. \tag{2}$$

This frames the general problem, but without further specialization, the tools for analysis are limited. Guerriero, Pozdnyakov, Glaz, and Willett [3] considered the case when B is large and used the renewal theorem approximation $\tau_E \sim B/\mathbf{E}(c(X_1))$ to make some progress, but here we are specifically concerned with the case for which the budget B is reasonably small. It is precisely for applications with a strongly binding budget that exact formulas are most useful.

3. BINOMIAL-PLUS-PASSIVE WALK

To have any hope for exact formulas, one needs more detailed information on distribution of the X_i, the cost function c, and the decision boundaries. The simplest nontrivial case begins with a trinomial model for the X_i that we parameterize as $P(X_i = 1) = p, P(X_i = 0) = r$, and $P(X_i = -1) = q$ with nonnegative p, r, and q such that $p + r + q = 1$. We then take the simplest possibilities for the decision boundaries; for the upper boundary $\mathcal{U}$, we take a constant U, and for the lower boundary $\mathcal{L}$, we take the constant $-L$, where $U > 0$ and $L > 0$ are integers.

The choice of a useful cost function is more subtle since the cost function must capture the benefit of the "nonresponse" possibility in the binary-plus-passive protocols. Still, only special choices are likely to yield tractable formulas. These considerations lead us to introduce an integer K and to consider the cost function defined by

$$c(x) = |x| + \delta(x)/K, \tag{3}$$

where $\delta(0) = 1$ and $\delta(x) = 0$ if $x \neq 0$. In other words, our cost for the binary transmission of 1 or -1 from a distributed sensor is a "unit," and the cost of the passive transmission from a distributed sensor (i.e., a nonresponse to a query) is just a $1/K$ fraction of a "unit."

The decision time (1) is now more explicitly given by

$$\tau_D = \min\{n : S_n \geq U \text{ or } S_n \leq -L\}, \tag{4}$$

and the budget exhaustion time is given by (3) and

$$\tau_E = \min\{n : c(X_1) + c(X_2) + \cdots + c(X_n) \geq B\}. \tag{5}$$

Here one should note that the total cost C_{τ_E} at the time the budget is exceeded can take on the any of the values $B, B + 1/K, \ldots, B + (K-1)/K$, but an "overshoot" of the budget is only possible if the sensor response at time τ_E was an active response.

There are several probabilities that can inform us about the design of a binary-plus-passive distributed sensor network. We are particularly interest in

$$\mathbf{P}(\tau_D \leq \tau_E), \tag{6}$$

the probability that we make a decision before we overrun our budget. We are also interested in this event together with the kind of decision that we make:

$$\mathbf{P}(\tau_D \leq \tau_E \text{ and } S_{\tau_D} = U) \quad \text{and} \quad \mathbf{P}(\tau_D \leq \tau_E \text{ and } S_{\tau_D} = -L). \tag{7}$$

Moreover, it may be useful sometimes to know just where the evidence stands when no decision has been made and yet the budget is exhausted; this is given by

$$\mathbf{P}(\tau_E < \tau_D \text{ and } S_{\tau_E} = x). \tag{8}$$

4. COMPUTATIONAL FORMULAS FOR PROBABILITIES OF INTEREST

Formulas for these probabilities of interest for the binary-passive distributed sensor model can be given with help from some related binary variables for which analogous probabilities are either well known or easily found. To describe these variables, first note that a $\{-1, 0, 1\}$ trinomial process can be associated with a binomial process simply by deleting zeros and the times at which they occur. This process of "casting out zeros" distorts the values (and distribution) of hitting times, but it does so in way that still permits useful calculations. To make this explicit, we first introduce a new i.i.d. sequence $\{X_i^* : i = 1, 2, \ldots\}$ with

$$\mathbf{P}(X_i^* = 1) = p_* \quad \text{and} \quad \mathbf{P}(X_i^* = -1) = q_*,$$

where

$$p_* = p/(p+q) \quad \text{and} \quad q_* = q/(p+q),$$

and we consider the new binomial random walk $S_n^* = X_1^* + X_2^* + \cdots + X_n^*$ together with a corresponding "decision time"

$$\tau_D^* = \min\{n : S_n^* \geq U \text{ or } S_n^* \leq -L\}. \tag{9}$$

Finally, it will be useful in our analysis to consider the number ν of active responses observed up to and including time τ_E. Under our binary-passive protocol, this is simply

$$\nu = \sum_{i=1}^{\tau_E} |X_i|. \tag{10}$$

We now have a theorem that tells shows how the basic probability results for τ_D, τ_E, S_{τ_D}, and S_{τ_E} can be expressed in terms of the more easily analyzed (or well-known) quantities ν, τ_D^*, τ_E^*, $S_{\tau_D}^*$, and $S_{\tau_E}^*$. In a later section we illustrate these computational relations with a numerical example.

THEOREM 1: *For each* $0 \le n \le B$, *one has*

$$\mathbf{P}(\nu = n) = \binom{(B-n)K+n}{n}(1-r)^n r^{(B-n)K} + \sum_{i=1}^{K-1} \binom{n-1+(B-n)K+i}{n-1}(1-r)^n r^{(B-n)K+i}.$$

Moreover, we have four identities that relate the "unstarred variables" to ν *and the simpler "starred variables":*

$$\mathbf{P}(\tau_D > \tau_E) = \sum_{n=0}^{B} \mathbf{P}(\tau_D^* > n)\,\mathbf{P}(\nu = n), \tag{11}$$

$$\mathbf{P}(\tau_D \le \tau_E, S_{\tau_D} = U) = \sum_{n=0}^{B} \mathbf{P}(\tau_D^* \le n, S_{\tau_D^*}^* = U)\,\mathbf{P}(\nu = n),$$

$$\mathbf{P}(\tau_D \le \tau_E, S_{\tau_D} = -L) = \sum_{n=0}^{B} \mathbf{P}(\tau_D^* \le n, S_{\tau_D^*}^* = -L)\,\mathbf{P}(\nu = n),$$

$$\mathbf{P}(\tau_D > \tau_E, S_{\tau_E} = x) = \sum_{n=0}^{B} \mathbf{P}(\tau_D^* > n, S_n^* = x)\,\mathbf{P}(\nu = n),$$

where $-L < x < U$.

Comment: One should note that given the first formula, all of the terms on the right side of the subsequent formulas can be readily computed since all the "starred" variables refer to the standard biased random walk $\{S_k^* : k = 0, 1, \ldots\}$ for which formulas (or methods) for all of the required probabilities are wellknown.

PROOF OF THEOREM 1: We will first prove the identity (11) and then derive the formula for $P(\nu = n)$. The proofs of the remaining formulas are similar and can be safely

omitted. To begin, we note that $0 \le \nu \le B$, so we have

$$\mathbf{P}(\tau_D > \tau_E) = \sum_{n=0}^{B} \mathbf{P}(\tau_D > \tau_E, \nu = n).$$

Next, we let $Q(n,j)$ denote number of paths $(S_1^*, S_2^*, \ldots, S_n^*)$ of length n of the affiliated binomial random walk such that the following hold:

1. The path $(S_1^*, S_2^*, \ldots, S_n^*)$ never hits $-L$ or U.
2. Exactly j of the summands of S_n^* are equal to $+1$.

We then have

$$\mathbf{P}(\tau_D^* > n) = \sum_{j=0}^{n} Q(n,j) p_*^j q_*^{n-j}.$$

We then consider padding the trajectory of the affiliated random walk of length n with k zero-valued summands.

The idea here is that each zero-valued summand has a cost of $1/K$ and we want these added summands to bring us precisely to the point where the budget is exhausted. Specifically, we choose k so that we have $C_{n+k-1} < B \le C_{n+k}$, and because of the possibility of overshooting, this gives us a range of values for k that is given by $(B-n)K \le k \le (B-n)K + K - 1$. This padding gives us a trajectory for a trinomial walk $(S_1, S_2, \ldots, S_{n+k})$ with the following properties:

1. $(S_1, S_2, \ldots, S_{n+k})$ stays inside the "no decision" interval $(-L, U)$.
2. S_{n+k} has j summands equal to $+1$.
3. S_{n+k} has $n-1$ summands equal to -1.
4. S_{n+k} has k summands equal to zero.

Here one should note that number $P(n,k)$ of ways of padding the affiliated walk with zeros depends *only* on n and k because adding such zeros does not change boundary crossing events. We then have the representation

$$\mathbf{P}(\nu = n) = \sum_{k=(B-n)K}^{(B-n)K+K-1} P(n,k)(p+q)^n r^k,$$

and for the moment, we do not need the exact values of the $P(n,k)$, although these will be found shortly. The proof of the identity (11) is now a calculation:

$$\begin{aligned}\mathbf{P}(\tau_D > \tau_E, \nu = n) &= \sum_{k=(B-n)K}^{(B-n)K+K-1} \sum_{j=0}^{n} Q(n,j) P(n,k), p^j q^{n-j} r^k \\ &= \sum_{k=(B-n)K}^{(B-n)K+K-1} P(n,k)(p+q)^n r^k \sum_{j=0}^{n} Q(n,j) p_*^j q_*^{n-j}\end{aligned}$$

$$= \sum_{k=(B-n)K}^{(B-n)K+K-1} P(n,k)(p+q)^n r^k \mathbf{P}(\tau_D^* > n)$$

$$= \mathbf{P}(\nu = n)\mathbf{P}(\tau_D^* > n).$$

Now, all we need to do now is to derive the distribution of ν. First, note that because of the possibility of going over the budget, we have the disjoint union

$$\Big\{\nu = n\Big\} = \Big\{\nu = n, C_{\tau_E} = B\Big\} \cup \Big\{\nu = n, C_{\tau_E} = B + 1/K\Big\} \cup \cdots$$
$$\cdots \cup \Big\{\nu = n, C_{\tau_E} = B + (K-1)/K\Big\}.$$

In the first event, there is no overshoot, and we calculate its probability separately:

$$\begin{aligned}\mathbf{P}(\nu = n, C_{\tau_E} = B) &= \mathbf{P}(|\{i \le \tau_E : |X_i| = 1\}| = n \\ &\qquad \text{and } |\{i \le \tau_E : |X_i| = 0\}| = (B-n)K) \\ &= \binom{(B-n)K+n}{n}(1-r)^n r^{(B-n)K}.\end{aligned}$$

In the other events, there is some overshoot, and one should note that overshooting is possible only if the last summand was $+1$ or -1. Thus, for $i = 1, 2, \ldots, K-1$, we see that $\mathbf{P}(\nu = n, C_{\tau_E} = B + i/K)$ is given by

$$(1-r)\binom{n-1+(B-n)K+i}{n-1}(1-r)^{n-1} r^{(B-n)K+i} \tag{12}$$

because out of the first $n-1+(B-n)K+i$ summands, exactly $(B-n)K+i$ are zero. Finally, for completeness, we note that when $n = 0$, overshooting is impossible, but, pleasantly enough, the formula still applies since, by convention, one has $\binom{j}{-1} = 0$. ■

Remark 1: What was actually shown in the first half of the proof is that because of linear boundaries, we have

$$\mathbf{P}(\tau_D > \tau_E \mid \nu = n) = \mathbf{P}(\tau_D^* > n).$$

Although this certainly seems intuitive, one can see from the argument above that a rigorous justification does require work.

Remark 2: From the proof one sees that we have identified $P(n,k)$ as

$$P(n,k) = \begin{cases} \binom{n+k}{n} & \text{if } k = (B-n)K \\ \binom{n-1+k}{n-1} & \text{if } (B-n)K < k \le (B-n)K + K - 1. \end{cases}$$

5. SMALL B CALCULATIONS VERSUS LARGE B APPROXIMATIONS

The main benefit of Theorem 1 is that it expresses nonstandard ruin probabilities such as $\mathbf{P}(\tau_D \le \tau_E)$ in terms the distribution of τ_D^*, the hitting time for a Bernoulli random walk. This hitting time is very well understood. For example, Feller [2, p. 351] gives a classic textbook treatment via generating functions, and there is a useful survey of modern developments by Lengyel [4].

The bottom line is that for B of small, or moderate size, Theorem 1 combines with the classical calculations to answer most of what one might ask about the probability $\mathbf{P}(\tau_D \le \tau_E)$ and related quantities. For large B, there is an alternative approach to that takes advantage of natural approximations from a renewal theory.

5.1. Two Renewal Theory Approximations

Guerriero et al. [3] observed that the strong renewal theorem tells us that as $B \to \infty$, one has

$$\frac{\tau_E}{B/\mathbf{E}(c(X_1))} = \frac{\tau_E}{B/(p+q+r/K)} \to 1$$

with probability 1, and this suggests some natural approximations to $\mathbf{P}(\tau_D \le \tau_E)$.

The first — and most natural — approximation is based directly on original trinomial walk. It is given by the formula

$$\mathbf{P}(\tau_D \le \tau_E) \approx \mathbf{P}\left(\tau_D \le \text{Round}\left[\frac{B}{p+q+r/K}\right]\right), \tag{13}$$

which one gets by replacing τ_E by its renewal theory approximation.

A more refined approximation uses the affiliated binomial random walk. Specifically, if τ_E is well represented by its renewal theory approximation $B/(p+q+r/K)$, then ν is well approximated by $B/(p+q+r/K) \times (p+q)$, because the fraction of the summands X_i that are different from zero is approximately $p+q$. These considerations motivate a second approximation:

$$\mathbf{P}(\tau_D \le \tau_E) \approx \mathbf{P}\left(\tau_D^* \le \text{Round}\left[\frac{B}{p+q+r/K} \times (p+q)\right]\right). \tag{14}$$

Finally, we should note that the use of rounding in the approximations (13) and (14) is reasonable — but somewhat arbitrary. One could just as well replace Round with Floor or Ceiling. We will also take these variations into consideration when we compare the *trinomial approximation* (13) and *binomial approximation* (14) to the exact values of $\mathbf{P}(\tau_D \le \tau_E)$ computed with the help of Theorem 1.

5.2. Numerical Comparisons

For the renewal theory approximations to have a fighting chance, the budget B must be of at least moderate size, so we first consider cases $B = 20$. To specify the rest of the

TABLE 1. Probability Estimates of Decision before Exhaustion in Example 1

	$\mathbf{P}(\tau_D \leq \tau_E)$		
Exact	.0113587		
Approximation	Round	Floor	Ceiling
Trinomial (13)	.0138173	.0138173	.0130238
Binomial (14)	.00683594	.00683594	.0147705

TABLE 2. Probability Estimates of Decision before Exhaustion in Example 2

	$\mathbf{P}(\tau_D \leq \tau_E)$		
Exact	.0971227		
Approximation	Round	Floor	Ceiling
Trinomial (13)	.0930403	.0876269	.0930403
Binomial (14)	.0980835	.0703125	.0980835

model, we take the signal probabilities to be $p = 1/10$ and $q = 1/10$ (so $r = 8/10$), take the decision limits to be $U = 10$ and $L = 10$, and, finally, take the cost to send a passive response to be $1/K = 1/8$. We then have a table of comparisons of the exact value for $\mathbf{P}(\tau_D \leq \tau_E)$ with the six candidate approximations given by (13) and (14) together with the variations that substitute "Floor" and "Ceiling" for "Round." (see Table 1.)

For the second example, we decrease B just a little to 17. We then take $p = 1/5$, $q = 1/5$, $r = 3/5$, $U = 8$, $L = 8$, and $K = 20$ to complete the model. (see Table 2.)

The first observation to be drawn from Tables 1 and 2 is that for such a moderate B, the approximation errors are relatively large. Moreover, there is no apparent way to try to make the renewal theory approximations much better through some "continuity correction." One needs to move the renewal theory approximations for τ_E to an integer value, and there does not appear to be any dominant choice among the three obvious alternatives. The absence of a good approximation (at least within the class of renewal theory alternatives) helps to underscore the benefit of computational formulas such as those provided by Theorem 1. In addition to its direct benefits, it provides a calibration on the accuracy of earlier approximations.

6. CONCLUDING REMARKS: DEALING WITH MORE REALISTIC BOUNDARIES

Here we have restricted attention to flat decision boundaries, but in the decision problems of most importance in sequential analysis, the decision boundaries are curved.

The assumption of level boundaries was essential to the effectiveness of the affiliated random walk, which set the stage for the conditioning used in Theorem 1.

At this moment, it is not at all clear how one would calculate—or even approximate—the analog of $\mathbf{P}(\tau_D \leq \tau_E)$ for parabolic boundaries or for boundaries given by sloping lines. In such cases, formulas of the style developed here seem highly unlikely.

As an intermediate step, one can consider multistage a sequential procedure with constant boundaries within each stage. For such problems, Theorem 1 provides useful guidance, and in many applied contexts, it is reasonable to expect that such multistage decision designs might offer effective approximations for decision problems with curved boundaries.

Still, in the absence of any exact theoretical results (or even numerical algorithms) for the corresponding curved boundary problems, it is not easy to say how one would evaluate the suggested multistage approximations. In the end, any approximations for the curved boundary problem would have to be evaluated by engineering judgement and experimentation. Theorem 1 at least provides a starting point.

References

1. Chong, C.-X. & Kumar, S.P. (2003). Sensor networks: Evolution, opportunities, and challenges. *Proceedings of the IEEE* 91, 1247–1256.
2. Feller, W. (1968). *An Introduction to probability theory and its applications*, Vol. 1, 3rd ed. New York: Wiley.
3. Guerriero, M., Pozdnyakov, V., Glaz, J., & Willett, P. (2010). A repeated significance test with applications to sequential detection in sensor networks. *IEEE Transactions on Signal Processing* 58: 3426–3435.
4. Lengyel, T. (to appear). Gambler's ruin & winning a series by m games. *Annals of the Institute of Statistical Mathematics* doi: 10.1007/s10463-008-0214-0.